"Where the data are coming from?" Ethics, crowdsourcing and traceability for Big Data in Human Language Technology

Gilles Adda, Laurent Besacier, Alain Couillault, Karen Fort, Joseph Mariani, Hugues de Mazancourt

"Where the data are coming from?" Ethics, crowdsourcing and traceability for Big Data in Human Language Technology

Gilles Adda*
LIMSI and IMMI-CNRS
Rue John von Neumann
91405 Orsay cedex France
adda@immi-labs.org

Laurent Besacier
Laboratoire d'Informatique de Grenoble
UJF BP53
38041 Grenoble Cedex 9, France
laurent.besacier@imag.fr

Alain Couillault
Université de La Rochelle/L3I
Av. Michel Crépeau
17042 La Rochelle Cedex 01, France
alain.couillault@univ-lr.fr

Karën Fort
Université de Lorraine/LORIA
54500 Vandœuvre-lès-Nancy, France
karen.fort@loria.fr

Joseph J. Mariani
LIMSI and IMMI-CNRS
Rue John von Neumann
91405 Orsay cedex France
joseph.mariani@limsi.fr

Hugues de Mazancourt
Eptica-Lingway
hugues.de-mazancourt@eptica.com

ABSTRACT

Based on the experience gained on the observation of the corpora developement in HLT, the authors want to warn the Big Data community about some recent usage of human computation. For instance, the growing use in the HLT community of crowdsourcing methods, and especially of microworking retributed crowsourcing platforms, lead to many ethical, economical and juridical concerns. The authors want also to foster some behaviours, especially concerning traceability, implemented in the form of a charter, the *Ethics and Big Data Charter*.

Keywords

Ethics, crowdsourcing, traceability

1. INTRODUCTION

At the turn of the 90s, Human Language Technology (HLT) changed drastically, with the developpement of probabilistic machine learning methods and the establishment of the evaluation paradigm. One crucial result implied by these changes is the development of corpora of growing sizes and of varying natures. From speech corpora of few hours and text corpora of few millions of words, we developed our models from Multimedia, Multimodal, Multilingual (3M) data containing for instance thousands of hours of video and terawords text corpora coming from social media, newspapers, subtitles, transcriptions, translations, etc.

To be able to manage the constitution and the distribution of these corpora, specific agencies have been developed within the HLT community, such as LDC[1] or ELDA.[2] But at a certain point, the growth of the needed corpora along the famous Big Data 3Vs[3] axes, goes beyond the classical well-controlled developement of corpora, relying on a meticulous and safe selection of data, and retributed employees to annotate them. We did see more and more corpora design involving the use of crowdsourcing, and especially the use of microworking, monetary compensated work platform such as Amazon Mechanical Turk. The authors, as academic and industrial HLT actors, wrote numerous positions papers to warn the community on ethical, economical and juridical concerns about MTurk [1, 7, 2, 8]. From these statements, we found that retributed crowdsourcing was only one (important) symptom of way of doing in the HLT community for the constitution and annotation of corpora.

Some of the authors have designed the *Ethics and Big Data Charter* [4] in collaboration with representatives from interest groups, private companies and academic organizations, including ELDA, ATALA,[4] AFCP[5] and APROGED.[6] The purpose of this Charter is to provide resources developers, not only in the HLT community, but more broadly in the Big Data community, with a precise framework to document their resources and ensure their traceability and transparency.

*Alphabetic order

[1]Linguistic Data Consortium, `http://www.ldc.upenn.edu/`
[2]Evaluations and Language resources Distribution Agency, `http://www.elda.org/`
[3]Volume, Velocity, Variety
[4]Association pour le Traitement Automatique des Langues `http://www.atala.org`
[5]Association Française de Communication Parlée `http://www.afcp-parole.org`
[6]Association de la Maîtrise et de la Valorisation des contenus `http://www.aproged.org`

MASTODONS Paris, France

2. TOWARDS TRACEABILITY: THE ETHICS AND BIG DATA CHARTER

To adopt an ethical behavior in developing, funding, using or promoting language resources is first and above all a matter of choice: for the provider, deciding which approach to adopt – crowdsourcing or not –, or which platform to request on, or the level of remuneration of the workers; for the funding agency, choosing which project to fund; for users, choosing which resource to use or acquire.

2.1 Contents of the Charter

The *Ethics and Big Data Charter* is provided as a form to be filled in by the dataset provider. It is split into three major sections: *traceability*, *intellectual property* and *specific legislation*, preceded by a short identification section containing the names of the resource, the contact and responsible persons and a short description of the data set. We describe below some of the important points listed in the charter.

2.1.1 Traceability

Traceability is key to our purpose of putting forward ethical issues. The traceability part of the charter allows to precise the relationship between the resource provider and the workers involved in developing the resource, including legal bounding, workers skills, selection criteria.
Specific focus is put on personal data, i.e. data such as voice or video recording, which can provide a means to identify a person directly or indirectly. The Charter requires to precise if and how the data is de-identified, and if and how the individuals were informed of the purpose of the data collection. Quality assurance is another major aspect of traceability addressed by the charter, as it requires to document the quality assurance strategy, so that the user of the data set is fully informed on the level of quality s/he can expect.

2.1.2 License and copyright

Thanks to a great deal of effort accomplished in the definition of – mainly open source – license schemes, it has become common practice to attach a license to a data set. The License and Copyright section of the Charter goes beyond this and puts the focus on questions which may be disregarded, like ensuring that the legal or moral copyrights of the persons who worked on compiling, enriching or transforming the data are respected.

2.1.3 Specific legal requirements

A third section of the *Ethics and Big Data Charter* deals with legal requirements that may arise from certain properties of the data set. For example, a country may have issued specific laws regarding the storing, use and/or dissemination of personal data. The Charter serves as a reminder for checking if such requirements exist.

2.2 Availability

The *Ethics and Big Data Charter* is available on-line.[7] Examples of charters are also provided.

3. CONCLUSION AND PERSPECTIVES

In [5], we showed that even the most widely referred to language resources present documentation lacks, in particular regarding the persons who produced the work and which external resources have been used. We assume that this also applies for many corpora used in the Big Data community.

The reality we are facing now in the call for projects of all the national and international funding research agencies is a growing interest for Big Data, and a foreseeable burst in the number of Big Data related projects, with the use, for instance, of personal data from Twitter or Facebook, the development of crowdsourcing and so on. But we also know that for these data, uncertainties about privacy, the way they have been acquired, who has really worked, is likely to be of an order of magnitude more important than for the present resources.

Other initiatives and articles have tackled this subject of Big data and Ethics (see for instance [6, 3, 9]), most of them dealing with the problem of privacy and surveillance.

For these reasons, we think it is crucial to gather all the initiatives such as the *Ethics and Big Data Charter* which aims at promoting ethics and traceability in resources, in order to propose, at the international level, a way to limit the risks for all the actors (funding agencies, research laboratories, private companies) of the data added value chain, regarding the use of Big Data and bring the benefits of the *Ethics and Big Data Charter* to current projects.

[7] `http://wiki.ethique-big-data.org`

4. REFERENCES

[1] G. Adda and J. Mariani. Language resources and amazon mechanical turk: legal, ethical and other issues. In *LISLR 2010, "Legal Issues for Sharing Language Resources workshop", LREC 2010*, Valletta, Malta, May 2010.

[2] G. Adda, J. Mariani, L. Besacier, and H. Gelas. *Crowdsourcing for Speech Processing*, chapter Economic and Ethical Background of Crowdsourcing for Speech, pages 303–334. Wiley, 2013.

[3] D. Boyd and K. Crawford. Critical questions for big data. 5(15):662–679, May 2012.

[4] A. Couillault and K. Fort. Charte Éthique et Big Data : parce que mon corpus le vaut bien ! In *Linguistique, Langues et Parole : Statuts, Usages et Mésusages*, Strasburg, France, July 2013. 4 pages.

[5] A. Couillault, K. Fort, G. Adda, and H. De Mazancourt. Evaluating Corpora Documentation with regards to the Ethics and Big Data Charter. In *International Conference on Language Resources and Evaluation (LREC)*, Reykjavik,Iceland, May 2014.

[6] K. Davis. *Ethics of Big Data: Balancing Risk and Innovation*. O'Reilly Media, Inc., 2012.

[7] K. Fort, G. Adda, and K. B. Cohen. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics (editorial)*, 37(2), 2011.

[8] K. Fort, G. Adda, B. Sagot, J. Mariani, and A. Couillault. *HLT Challenges for Computer Science and Linguistics*, chapter Crowdsourcing for Language Resource Development: Criticisms about Amazon Mechanical Turk Overpowering Use. Springer, 2014.

[9] N. M. Richards and J. H. King. Big Data Ethics. In *Wake Forest Law Review*, May 2014. Available at SSRN: `http://ssrn.com/abstract=2384174`.